Approximation of a Maximum-Submodular-Coverage problem involving spectral functions, with application to Experimental Design

Guillaume Sagnol[*]
Zuse Institut Berlin (ZIB), Takustr. 7, 14195 Berlin, Germany
sagnol@zib.de

Abstract

We study a family of combinatorial optimization problems defined by a parameter $p \in [0,1]$, which involves spectral functions applied to positive semidefinite matrices, and has some application in the theory of optimal experimental design. This family of problems tends to a generalization of the classical maximum coverage problem as p goes to 0, and to a trivial instance of the knapsack problem as p goes to 1.

In this article, we establish a matrix inequality which shows that the objective function is submodular for all $p \in [0,1]$, from which it follows that the greedy approach, which has often been used for this problem, always gives a design within $1-1/e$ of the optimum. We next study the design found by rounding the solution of the continuous relaxed problem, an approach which has been applied by several authors. We prove an inequality which generalizes a classical result from the theory of optimal designs, and allows us to give a rounding procedure with an approximation factor which tends to 1 as p goes to 1.

Keyword Maximum Coverage, Optimal Design of Experiments, Kiefer's $p-$criterion, Polynomial-time approximability, Rounding algorithms, Submodularity, Matrix inequalities.

1 Introduction

This work is motivated by a generalization of the classical maximum coverage problem which arises in the study of optimal experimental designs. This problem may be formally defined as follows: given s positive semidefinite matrices $M_1, \ldots, M_s$ of the same size and an integer $N < s$, solve:

$$\begin{aligned} \max_{I \subset [s]} \quad & \operatorname{rank}\Big(\sum_{i \in I} M_i\Big) \qquad (P_0) \\ \text{s.t.} \quad & \operatorname{card}(I) \leq N, \end{aligned}$$

[*] Parts of this work were done when the author was with INRIA Saclay Île de France & CMAP, École Polytechnique, being supported by Orange Labs through the research contract CRE EB 257676 with INRIA.

where we use the standard notation $[s] := \{1, \dots, s\}$ and $\mathrm{card}(S)$ denotes the cardinality of S. When each M_i is diagonal, it is easy to see that Problem (P_0) is equivalent to a max-coverage instance, by defining the sets $S_i = \{k : (M_i)_{k,k} > 0\}$, so that the rank in the objective of Problem (P_0) is equal to $\mathrm{card}\big(\cup_{i\in I} S_i\big)$.

A more general class of problems arising in the study of optimal experimental designs is obtained by considering a *deformation* of the rank which is defined through a spectral function. Given $p \in [0, 1]$, solve:

$$\begin{aligned} \max_{\boldsymbol{n}\in\mathbb{N}^s} \quad & \varphi_p(\boldsymbol{n}) \\ \text{s. t.} \quad & \sum_{i\in[s]} n_i \leq N, \end{aligned} \tag{P_p}$$

where $\varphi_p(\boldsymbol{n})$ is the sum of the eigenvalues of $\sum_{i\in[s]} n_i M_i$ raised to the exponent p: if the eigenvalues of the positive semidefinite matrix $\sum_{i\in[s]} n_i M_i$ are $\lambda_1, \dots, \lambda_m$ (counted with multiplicities), $\varphi_p(\boldsymbol{n})$ is defined by

$$\varphi_p(\boldsymbol{n}) = \mathrm{trace}\Big(\sum_{i\in[s]} n_i M_i\Big)^p = \sum_{k=1}^m \lambda_k^p.$$

We shall see that Problem (P_0) is the limit of Problem (P_p) as $p \to 0^+$ indeed. On the other hand, the limit of Problem (P_p) as $p \to 1$ is a knapsack problem (in fact, it is the trivial instance in which the i^{th} item has weight 1 and utility $u_i = \mathrm{trace}\, M_i$). Note that a matrix M_i may be chosen n_i times in Problem (P_p), while choosing a matrix more than once in Problem (P_0) cannot increase the rank. Therefore we also define the binary variant of Problem (P_p):

$$\max_{\boldsymbol{n}} \left\{ \varphi_p(\boldsymbol{n}) : \ \boldsymbol{n} \in \{0,1\}^s, \ \sum_{i\in[s]} n_i \leq N \right\} \tag{P_p^{bin}}$$

We shall also consider the case in which the selection of the i^{th} matrix costs c_i, and a total budget B is allowed. This is the budgeted version of the problem:

$$\max_{\boldsymbol{n}} \left\{ \varphi_p(\boldsymbol{n}) : \ \boldsymbol{n} \in \mathbb{N}^s, \ \sum_{i\in[s]} c_i n_i \leq B \right\} \tag{P_p^{bdg}}$$

Throughout this article, we use the term *design* for the variable $\boldsymbol{n} = (n_1, \dots, n_s) \in \mathbb{N}^s$. We say that $\boldsymbol{n}$ is a $N-$*replicated design* if it is feasible for Problem (P_p), a $N-$*binary design* if $\boldsymbol{n}$ is feasible for Problem (P_p^{bin}), and a $B-$*budgeted design* when it satisfies the constraints of (P_p^{bdg}).

1.1 Motivation: optimal experimental design

The theory of *optimal design of experiments* plays a central role in statistics. It studies how to best select experiments in order to estimate a set of parameters. Under classical assumptions, the best linear unbiased estimator is given by least square theory, and lies within confidence ellipsoids which are described by a positive semidefinite matrix depending only on the selected experiments. The *optimal design of experiments* aims at selecting the experiments in order to make these confidence ellipsoids as small as possible, which leads to more accurate estimators.

A common approach consists in minimizing a scalar function measuring these ellipsoids, where the function is taken from the class of Φ_p-information functions proposed by Kiefer [Kie75]. This leads to a combinatorial optimization problem (decide how many times each experiment should be performed) involving a spectral function which is applied to the information matrix of the experiments. For $p \in]0,1]$, the Kiefer's Φ_p-optimal design problem is equivalent to Problem (P_p) (up to the exponent $1/p$ in the objective function).

In fact, little attention has been given to the combinatorial aspects of Problem (P_p) in the optimal experimental design literature. The reason is that there is a natural relaxation of the problem which is much more tractable and usually yields very good results: instead of determining the exact number of times n_i that each experiment will be selected, the optimization is done over the fractions $w_i = n_i/N \in [0,1]$, which reduces the problem to the maximization of a concave function over a convex set (this is the theory of *approximate optimal designs*). For the common case, in which the number N of experiments to perform is large and $N > s$ (where s is the number of available experiments), this approach is justified by a result of Pukelsheim and Rieder [PR92], who give a rounding procedure to transform an optimal approximate design $\boldsymbol{w}^*$ into an $N-$replicated design $\boldsymbol{n} = (n_1, \ldots, n_s)$ which approximates the optimum of the Kiefer's Φ_p-optimal design problem within a factor $1 - \frac{s}{N}$.

The present developments were motivated by a joint work with Bouhtou and Gaubert [BGS08, SGB10] on the application of optimal experimental design methods to the identification of the traffic in an Internet backbone. This problem describes an *underinstrumented situation*, in which a small number $N < s$ of experiments should be selected. In this case, the combinatorial aspects of Problem (P_p) become crucial. A similar problem was studied by Song, Qiu and Zhang [SQZ06], who proposed to use a greedy algorithm to approximate the solution of Problem (P_p). In this paper, we give an approximation bound which justifies this approach. Another question addressed in this manuscript is whether it is appropriate to take roundings of (continuous) approximate designs in the underinstrumented situation (recall that this is the common approach when dealing with experimental design problems in the *overinstrumented* case, where the number N of experiments is large when compared to s).

Appendix A is devoted to the application to the theory of optimal experimental designs; we explain how a statistical problem (choose which experiments to conduct in order to estimate a set of parameters) leads to the study of Problem (P_p), with a particular focus to the *underinstrumented* situation described above. For more details on the subject, the reader is referred to the monographs of Fedorov [Fed72] and Pukelsheim [Puk93].

1.2 Organisation and contribution of this article

The objective of this article is to study some approximation algorithms for the class of problems $(P_p)_{p \in [0,1]}$. Several results presented in this article were already announced in the companion papers [BGS08, BGS10], without the proofs. This paper provides all the proofs of the results of [BGS10] and gives new results for the rounding algorithms. We shall now present the contribution and the organisation of this article.

In Section 2, we establish a matrix inequality (Proposition 2.3) which shows that a class

Algorithm	Approximation factor for Problem (P_p)	Reference
Greedy	$1-e^{-1} \qquad (\text{or } 1-(1-\frac{1}{N})^N)$	2.6 ([NWF78])
Any $N-$replicated design $\boldsymbol{n}$ (posterior bound)	$\frac{1}{N}\sum_{i=1}^{s} n_i^p (w_i^*)^{1-p}$	3.3
Rounding 3.1 (prior bound)	$\begin{cases} \left(\frac{N}{s}\right)^{1-p} & \text{if } \left(\frac{N}{s}\right)^{1-p} \leq \frac{1}{2-p}; \\ 1-\frac{s}{N}(1-p)\left(\frac{1}{2-p}\right)^{\frac{2-p}{1-p}} & \text{Otherwise} \end{cases}$	3.8
Apportionment rounding	$(1-\frac{s}{N})^p \qquad \text{if } N \geq s$	[PR92]

Algorithm	Approximation factor for Problem (P_p^{bin})	Reference
Greedy	$1-e^{-1} \qquad (\text{or } 1-(1-\frac{1}{N})^N)$	2.6 ([NWF78])
Any $N-$binary design $\boldsymbol{n}$ (posterior bound)	$\frac{1}{N}\sum_{i=1}^{s} n_i (w_i^*)^{1-p}$	3.1
Keep the N largest coord. of $\boldsymbol{w}^*$ (prior bound)	$\left(\frac{N}{s}\right)^{1-p} \qquad \text{if } p \leq 1-\frac{\ln N}{\ln s}$	3.7

Algorithm	Approximation factor for Problem (P_p^{bdg})	Reference
Adapted Greedy	$1-e^{-\beta} \simeq 0.35$ (where $e^{\beta}=2-\beta$)	2.8([Wol82])
Greedy+triples enumeration	$1-e^{-1}$	2.8([Svi04])
Any $B-$budgeted design $\boldsymbol{n}$ (posterior bound)	$\frac{1}{B}\sum_{i=1}^{N} c_i n_i^p (w_i^*)^{1-p}$	3.5

Table 1: Summary of the approximation bounds obtained in this paper, as well as the bound of Pukelsheim and Rieder [PR92]. The column "Reference" indicates the number of the theorem, proposition or remark where the bound is proved (a citation in parenthesis means a direct application of a result of the cited paper, which is possible thanks to the submodularity of φ_p proved in Corollary 2.5). In the table, *posterior* denotes a bound which depends on the continuous solution $\boldsymbol{w}^*$ of the relaxed problem, while a *prior bound* depends only on the parameters of the problem.

of spectral functions is submodular (Corollary 2.4). As a particular case of the latter result, the objective function of Problem (P_p) is submodular for all $p \in [0,1]$. The submodularity of this class of spectral functions is an original contribution of this article for $0 < p < 1$, however a particular case of this result was announced –without a proof– in the companion paper on the telecom application [BGS08]. In the limit case $p = 0$, we obtain two functions which were already known to be submodular (the rank and the log of determinant of a sum of matrices).

Due to a celebrated result of Nemhauser, Wolsey and Fisher [NWF78], the submodularity of the criterion implies that the greedy approach, which has often been used for this problem, always gives a design within $1-e^{-1}$ of the optimum (Theorem 2.6). We point out that the submodularity of the determinant criterion was noticed earlier in the optimal experimental design literature, but under an alternative form [RS89]: Robertazzi and Schwartz showed that the determinant of the inverse of a sum of matrices is supermodular, and they used it to write an algorithm for the construction of approximate designs (i.e. without integer variables) which is based on the accelerated greedy algorithm of Minoux [Min78]. In contrast, the originality of the present paper is to show that a whole class of criteria satisfies the submodularity property, and to study the consequences in terms of approximability of a combinatorial optimization problem.

In Section 3, we investigate the legitimacy of using rounding algorithms to construct a $N-$replicated design $\boldsymbol{n} = (n_1, \ldots, n_s) \in \mathbb{N}^s$ or a N-binary design $\boldsymbol{n} \in \{0,1\}^s$ from an optimal approximate design $\boldsymbol{w}^*$, i.e. a solution of a continuous relaxation of Problem (P_p). We establish an inequality (Propositions 3.1 and 3.3) which bounds from below the approximation ratio of any integer design, by a function which depends on the continuous solution $\boldsymbol{w}^*$. Interestingly, this inequality generalizes a classical result from the theory of optimal designs (the upper bound on the weights of a D-optimal design [Puk80, HT09] is a particular case $(p = 0)$ of Proposition 3.1). The proof of this result is presented in Appendix B ; it relies on matrix inequalities and several properties of the differentiation of a scalar function applied to symmetric matrices. Then we point out that the latter lower bound can be maximized by an incremental algorithm which is well known in the resource allocation community (Algorithm 3.1), and we derive approximation bounds for Problems (P_p) and (P_p^{bin}) which do not depend on $\boldsymbol{w}^*$ (Theorems 3.7 and 3.8). For the problem with replicated designs (P_p), the approximation factor is an increasing function of p which tends to 1 as $p \to 1$. In many cases, the approximation guarantee for designs obtained by rounding is better than the greedy approximation factor $1 - e^{-1}$.

We have summarized in Table 1 the approximation results proved in this paper (this table also includes another known approximability result for Problem (P_p), the *efficient apportionment rounding* of Pukelsheim and Rieder [PR92]).

2 Submodularity and Greedy approach

In this section, we study the greedy algorithm for solving Problems (P_p) and (P_p^{bin}) through the submodularity of φ_p. We first recall a result presented in [BGS08], which states that the *rank optimization* problem is NP-hard, by a reduction from the *Maximum Coverage* problem. It follows that for all positive ε, there is no polynomial-time algorithm which approximates (P_0) by a factor of $1 - \frac{1}{e} + \varepsilon$ unless $P = NP$ (this has been proved by Feige for the Maximum Coverage problem [Fei98]). Nevertheless, we show that this bound is the worst possible ever, and that the greedy algorithm always attains it.

To this end, we show that a class of spectral functions (which includes the objective function of Problem (P_p)) is *nondecreasing submodular*. The maximization of submodular functions over a matroid has been extensively studied [NWF78, CC84, CCPV07, Von08, KST09], and we shall use known approximability results.

To study its approximability, we can think of Problem (P_p) as the maximization of a set function $\varphi_p' : 2^E \mapsto \mathbb{R}^+$. To this end, note that each design $\boldsymbol{n}$ can be seen as a subset of E, where E is a pool which contains N copies of each experiment (this allows us to deal with replicated designs, i.e. with experiments that are conducted several times; if replication is not allowed (Problem (P_p^{bin})), we simply set $E := [s]$). Now, if S is a subset of E corresponding to the design $\boldsymbol{n}$, we define $\varphi_p'(S) := \varphi_p(\boldsymbol{n})$. In the sequel, we identify the set function φ_p' with φ_p (i.e., we omit the *prime*).

We also point out that multiplicative approximation factors for the Φ_p-optimal problem cannot be considered when $p \leq 0$, since the criterion is identically 0 as long as the the information matrix is singular. For $p \leq 0$ indeed, the instances of the Φ_p-optimal problem where no feasible design lets $M_F(\boldsymbol{n})$ be of full rank have an optimal value of 0. For all the other instances, any polynomial-time algorithm with a positive approximation factor would necessarily return a design of full rank. Provided that $P \neq NP$, this would contradict the NP-hardness of *Set-Cover* (it is easy to see that *Set Cover* reduces to the problem of deciding whether there exists a set S of cardinal N such that $\sum_{i \in S} M_i$ has full rank for some diagonal matrices M_i, by a similar argument to the one given in the first paragraph of this article). Hence, we investigate approximation algorithms only in the case $p \in [0, 1]$.

2.1 A class of submodular spectral functions

In this section, we are going to show that a class of spectral functions is submodular. We recall that a real valued function $F : 2^E \to \mathbb{R}$, defined on every subset of E is called nondecreasing if for all subsets I and J of E, $I \subseteq J$ implies $F(I) \leq F(J)$. We also give the definition of a *submodular* function:

Definition 2.1 (Submodularity)**.** A real valued set function $F : 2^E \longrightarrow \mathbb{R}$ is *submodular* if it satisfies the following condition :

$$F(I) + F(J) \geq F(I \cup J) + F(I \cap J) \quad \text{for all} \quad I, J \subseteq E.$$

We next recall the definition of operator monotone functions. The latter are real valued functions applied to hermitian matrices: if $A = U \operatorname{Diag}(\lambda_1, \ldots, \lambda_m) U^*$ is a $m \times m$ hermitian matrix (where U is unitary and U^* is the conjugate of U), the matrix $f(A)$ is defined as $U \operatorname{Diag}(f(\lambda_1), \ldots, f(\lambda_m)) U^*$.

Definition 2.2 (Operator monotonicity)**.** A real valued function f is *operator monotone* on $\mathbb{R}_+$ (resp. $\mathbb{R}_+^*$) if for every pair of positive semidefinite (resp. positive definite) matrices A and B,

$$A \preceq B \Longrightarrow f(A) \preceq f(B).$$

We say that f is *operator antitone* if $-f$ is operator monotone.

The next proposition is a matrix inequality of independent interest; it will be useful to show that φ_p is submodular. Interestingly, it can be seen as an extension of the Ando-Zhan Theorem [AZ99], which reads as follows: *Let A, B be semidefinite positive matrices. For any unitarily invariant norm $\||\cdot\||$, and for every non-negative operator monotone function f on* $[0, \infty)$,

$$\||f(A + B)\|| \leq \||f(A) + f(B)\||.$$

Kosem [Kos06] asked whether it is possible to extend this inequality as follows:

$$\||f(A + B + C)\|| \leq \||f(A + B) + f(B + C) - f(C)\||,$$

and gave a counterexample involving the trace norm and the function $f(x) = \frac{x}{x+1}$. However, we show in next proposition that the previous inequality holds for the trace norm and every primitive f of an operator antitone function (in particular, for $f(x) = x^p,\ p \in]0,1]$). Note that the previous inequality is not true for any unitarily invariant norm and $f(x) = x^p$ either. It is easy to find counterexamples with the spectral radius norm.

Proposition 2.3. *Let f be a real function defined on $\mathbb{R}_+$ and differentiable on $\mathbb{R}_+^*$. If f' is operator antitone on $\mathbb{R}_+^*$, then for all triples (X, Y, Z) of $m \times m$ positive semidefinite matrices,*

$$\text{trace } f(X+Y+Z) + \text{trace } f(Z) \leq \text{trace } f(X+Z) + \text{trace } f(Y+Z). \tag{1}$$

Proof. Since the eigenvalues of a matrix are continuous functions of its entries, and since $\mathbb{S}_m^{++}$ is dense in $\mathbb{S}_m^+$, it suffices to establish the inequality when X, Y, and Z are positive definite. Let X be an arbitrary positive definite matrix. We consider the map:

$$\begin{aligned} \psi : \mathbb{S}_m^+ &\longrightarrow \mathbb{R} \\ T &\longmapsto \text{trace } f(X+T) - \text{trace } f(T). \end{aligned}$$

The inequality to be proved can be rewritten as

$$\psi(Y+Z) \leq \psi(Z).$$

We will prove this by showing that ψ is nonincreasing with respect to the Löwner ordering in the direction generated by any positive semidefinite matrix. To this end, we compute the Frechet derivative of ψ at $T \in \mathbb{S}_m^{++}$ in the direction of an arbitrary matrix $H \in \mathbb{S}_m^+$. By definition,

$$D\psi(T)(H) = \lim_{\epsilon \to 0} \frac{1}{\epsilon}\big(\psi(T+\epsilon H) - \psi(T)\big). \tag{2}$$

When f is an analytic function, $X \longmapsto \text{trace } f(X)$ is Frechet-differentiable, and an explicit form of the derivative is known (see [HP95, JB06]): $D\big(\text{trace } f(A)\big)(B) = \text{trace}\big(f'(A)B\big)$. Since f' is operator antitone on $\mathbb{R}_+^*$, a famous result of Löwner [Löw34] tells us (in particular) that f' is analytic at all points of the positive real axis, and the same holds for f. Provided that the matrix T is positive definite (and hence $X+T$), we have

$$D\psi(T)(H) = \text{trace}\Big(\ \big(f'(X+T) - f'(T)\big)H\Big).$$

By antitonicity of f' we know that the matrix $W = f'(X+T) - f'(T)$ is negative semidefinite. For a matrix $H \succeq 0$, we have therefore:

$$D\psi(T)(H) = \text{trace } (WH) \leq 0.$$

Consider now $h(s) := \psi(sY+Z)$. For all $s \in [0,1]$, we have

$$h'(s) = D\psi(sY+Z)(Y) \leq 0,$$

and so, $h(1) = \psi(Y+Z) \leq h(0) = \psi(Z)$, from which the desired inequality follows. □

Corollary 2.4. *Let $M_1,\ldots,M_s$ be $m\times m$ positive semidefinite matrices. If f satisfies the assumptions of Proposition 2.3, then the set function $F: 2^{[s]}\to\mathbb{R}$ defined by*

$$\forall I\subset[s],\ F(I)=\text{trace }\ f(\sum_{i\in I}M_i),$$

is submodular.

Proof. Let $I,J\subseteq 2^{[s]}$. We define

$$X=\sum_{i\in I\setminus J}M_i,\ Y=\sum_{i\in J\setminus I}M_i,\ Z=\sum_{i\in I\cap J}M_i.$$

It is easy to check that

$$\begin{aligned}F(I)&=\text{trace }f(X+Z),\\ F(J)&=\text{trace }f(Y+Z),\\ F(I\cap J)&=\text{trace }f(Z),\\ F(I\cup J)&=\text{trace }f(X+Y+Z).\end{aligned}$$

Hence, Proposition 2.3 proves the submodularity of F. □

A consequence of the previous result is that the objective function of Problem (P_p) is submodular. In the limit case $p\to 0^+$, we find two well-known submodular functions:

Corollary 2.5. *Let $M_1,...,M_s$ be $m\times m$ positive semidefinite matrices.*

(i) $\forall p\in]0,1], I\mapsto \text{trace}(\sum_{i\in I}M_i)^p$ *is submodular.*

(ii) $I\mapsto\text{rank}(\sum_{i\in I}M_i)$ *is submodular.*

If moreover every M_i is positive definite, then:

(iii) $I\mapsto\log\det(\sum_{i\in I}M_i)$ *is submodular.*

Proof. It is known that $x\mapsto x^q$ is operator antitone on $\mathbb{R}_+^*$ for all $q\in[-1,0[$. Therefore, the derivative of the function $x\mapsto x^p$ (which is px^{p-1}), is operator antitone on $\mathbb{R}_+^*$ for all $p\in]0,1[$. This proves the point (i) for $p\neq 1$. The case $p=1$ is trivial, by linearity of the trace.

The submodularity of the rank (ii) and of $\log\det$ (iii) are classic. Interestingly, they are obtained as the limit case of (i) as $p\to 0^+$. (For $\log\det$, we must consider the second term in the asymptotic development of $X\mapsto\text{trace }\ X^p$ as p tends to 0^+, cf. Equation (24)). □

2.2 Greedy approximation

We next present some consequences of the submodularity of φ_p for the approximability of Problem (P_p). Note that the results of this section hold in particular for $p=0$, and hence for the *rank maximization* problem (P_0). They also hold for $E=[s]$, i.e. for Problem (P_p^{bin}). We recall that the principle of the greedy algorithm is to start from $\mathcal{G}_0=\emptyset$ and to construct sequentially the sets

$$\mathcal{G}_{k+1}:=\mathcal{G}_k\cup\text{argmax}_{i\in E\setminus\mathcal{G}_k}\ \varphi_p(\mathcal{G}_k\cup\{i\}),$$

until $k=N$.

Theorem 2.6 (Approximability of Problem (P_p)). *Let $p \in [0,1]$. The greedy algorithm always yields a solution within a factor $1-\frac{1}{e}$ of the optimum of Problem (P_p).*

Proof. We know from Corollary 2.5 that for all $p \in [0,1]$, φ_p is submodular ($p=0$ corresponding to the rank maximization problem). In addition, the function φ_p is nondecreasing, because $X \longrightarrow X^p$ is a matrix monotone function for $p \in [0,1]$ (see e.g. [Zha02]) and $\varphi_p(\emptyset)=0$.

Nemhauser, Wolsey and Fisher [NWF78] proved the result of this theorem for any nondecreasing submodular function f satisfying $f(\emptyset) = 0$ which is maximized over a uniform matroid. Moreover when the maximal number of matrices which can be selected is N, this approximability ratio can be improved to $1-\big(1-1/N\big)^N$. □

Remark 2.7. One can obtain a better bound by considering the *total curvature* of a given instance, which is defined by:

$$c = \max_{i\in[s]} \quad 1-\frac{\varphi_p\big(E\big)-\varphi_p\big(E\setminus\{i\}\big)}{\varphi_p\big(\{i\}\big)} \in [0,1].$$

Conforti and Cornuejols [CC84] proved that the greedy algorithm always achieves a $\frac{1}{c}\big(1-(1-\frac{c}{N})^N\big)$-approximation factor for the maximization of an arbitrary nondecreasing submodular function with total curvature c. In particular, since φ_1 is additive it follows that the total curvature for $p=1$ is $c=0$, yielding an approximation factor of 1:

$$\lim_{c\to 0^+}\frac{1}{c}\big(1-(1-\frac{c}{N})^N\big)=1.$$

As a consequence, the greedy algorithm always gives the optimal solution of the problem. Note that Problem (P_1) is nothing but a *knapsack* problem, for which it is well known that the greedy algorithm is optimal if each available item has the same weight. However, it is not possible to give an upper bound on the total curvature c for other values of $p \in [0,1[$, and c has to be computed for each instance.

Remark 2.8. The problem of maximizing a nondecreasing submodular function subject to a budget constraint of the form $\sum_i c_i n_i \leq B$, where $c_i \geq 0$ is the cost for selecting the element i and B is the total allowed budget, has been studied by several authors. Wolsey presented an adapted greedy algorithm [Wol82] with a proven approximation guarantee of $1-e^{-\beta} \simeq 0.35$, where β is the unique root of the equation $e^x = 2-x$. More recently, Sviridenko [Svi04] showed that the budgeted submodular maximization problem was still $1-1/e-$approximable in polynomial time, with the help of an algorithm which associates the greedy with a partial enumeration of every solution of cardinality 3.

We have attained so far an approximation factor of $1-e^{-1}$ for all $p \in [0,1[$, while we have a guarantee of optimality of the greedy algorithm for $p=1$. This leaves a feeling of mathematical dissatisfaction, since intuitively the problem should be easy when p is very close to 1. In the next section we remedy to this problem, by giving a rounding algorithm with an approximation factor $F(p)$ which depends on p, and such that $p \mapsto F(p)$ is continuous, nondecreasing and $\lim_{p\to 1} F(p) = 1$.

3 Approximation by rounding algorithms

The optimal design problem has a natural continuous relaxation which is simply obtained by removing the integer constraint on the design variable $\boldsymbol{n}$, and has been extensively studied [Atw73, DPZ08, Yu10, Sag11]. As mentioned in the introduction, several authors proposed to solve this continuous relaxation and to round the solution to obtain a near-optimal discrete design. While this process is well understood when $N \geq s$, we are not aware of any bound justifying this technique in the underinstrumented situation $N < s$.

3.1 A continuous relaxation

The continuous relaxation of Problem (P_p) which we consider is obtained by replacing the integer variable $\boldsymbol{n} \in \mathbb{N}^s$ by a continuous variable $\boldsymbol{w}$ in Problem (22):

$$\max_{\substack{\boldsymbol{w}\ \in(\mathbb{R}_+)^s \\ \sum_k w_k \leq N}} \quad \Phi_p(M_F(\boldsymbol{w})) \tag{3}$$

Note that the criterion $\varphi_p(\boldsymbol{w})$ is raised to the power $1/p$ in Problem (3) (we have $\Phi_p(M_F(\boldsymbol{w})) = m^{-1/p}\varphi_p(\boldsymbol{w})^{1/p}$ for $p > 0$). The limit of Problem (3) as $p \to 0^+$ is hence the maximization of the determinant of $M_F(\boldsymbol{w})$ (cf. Equation (20)).

We assume without loss of generality that the matrix $M_F(\mathbf{1}) = \sum_{k=1}^s M_k$ is of full rank (where $\mathbf{1}$ denotes the vector of all ones). This ensures the existence of a vector $\boldsymbol{w}$ which is feasible for Problem (3), and such that $M_F(\boldsymbol{w})$ has full rank. If this is not the case ($r^* := \operatorname{rank}(M_F(\mathbf{1})) < m$), we define instead a projected version of the continuous relaxation: Let $U\Sigma U^T$ be a singular value decomposition of $M_F(\mathbf{1})$. We denote by U_{r^*} the matrix formed with the r^* leading singular vectors of $M_F(\mathbf{1})$, i.e. the r^* first columns of U. It can be seen that Problem (3) is equivalent to the problem with projected information matrices $\bar{M}_k := U_{r^*}^T M_k U_{r^*}$ (see Paragraph 7.3 in [Puk93]).

The functions $X \mapsto \log(\det(X))$ and $X \mapsto X^p$ ($p \in]0,1]$) are strictly concave on the interior of $\mathbb{S}_m^+$, so that the continuous relaxation (3) can be solved by interior-points technique or multiplicative algorithms [Atw73, DPZ08, Yu10, Sag11]. The strict concavity of the objective function indicates in addition that Problem (3) admits a unique solution if and only if

$$w_1M_1 + w_2M_2 + \ldots + w_sM_s = y_1M_1 + y_2M_2 + \ldots + y_sM_s \Rightarrow (w_1, \ldots, w_s) = (y_1, \ldots, y_s),$$

that is to say whenever the matrices M_i are linearly independent. In this paper, we focus on the rounding techniques only, and we assume that an optimal solution $\boldsymbol{w^*}$ of the relaxation (3) is already known. In the sequel, we also denote a discrete solution of Problem (P_p) by $\boldsymbol{n}^*$ and a binary solution of Problem (P_p^{bin}) by S^*. Note that we always have $\varphi_p(\boldsymbol{w^*}) \geq \varphi_p(\boldsymbol{n}^*) \geq \varphi_p(S^*)$.

3.2 Posterior bounds

In this section, we are going to bound from below the approximation ratio $\varphi_p(\boldsymbol{n})/\varphi_p(\boldsymbol{w^*})$ for an arbitrary discrete design $\boldsymbol{n}$, and we propose a rounding algorithm which maximizes this approximation factor. The lower bound depends on the continuous optimal variable $\boldsymbol{w^*}$, and hence we refer it as a *posterior* bound. We start with a result for binary designs ($\forall i \in [s], n_i \leq$

1), which we associate with a subset S of $[s]$ as in Section 2. The proof relies on several matrix inequalities and technical lemmas on the directional derivative of a scalar function applied to a symmetric matrix, and is therefore presented in Appendix B.

Proposition 3.1. *Let $p \in [0,1]$ and $\boldsymbol{w}^*$ be optimal for the continuous relaxation (3) of Problem (P_p). Then, for any subset S of $[s]$, the following inequality holds:*

$$\frac{1}{N}\sum_{i\in S}(w_i^*)^{1-p} \leq \frac{\varphi_p(S)}{\varphi_p(\boldsymbol{w}^*)}.$$

Remark 3.2. In this proposition and in the remaining of this article, we adopt the convention $0^0 = 0$.

We point out that this proposition includes as a special case a result of Pukelsheim [Puk80], already generalized by Harman and Trnovská [HT09], who obtained:

$$\frac{w_i^*}{N} \leq \frac{\operatorname{rank} M_i}{m},$$

i.e. the inequality of Proposition 3.1 for $p = 0$ and a singleton $S = \{i\}$. However the proof is completely different in our case. Note that there is no constraint of the form $w_i \leq 1$ in the continuous relaxation (3), although the previous proposition relates to binary designs $S \in [s]$. Proposition 3.1 suggests to select the N matrices with the largest coordinates w_i^* to obtain a candidate S for optimality of the binary problem (P_p^{bin}). We will give in the next section a *prior bound* (i.e., which does not depend on $\boldsymbol{w}^*$) for the efficiency of this rounded design.

We can also extend the previous proposition to the case of replicated designs $\boldsymbol{n} \in \mathbb{N}^s$ (note that the following proposition does not require the design $\boldsymbol{n}$ to satisfy $\sum_i n_i = N$):

Proposition 3.3. *Let $p \in [0,1]$ and $\boldsymbol{w}^*$ be optimal for the continuous relaxation (3) of Problem (P_p). Then, for any design $\boldsymbol{n} \in \mathbb{N}^s$, the following inequality holds:*

$$\frac{1}{N}\sum_{i\in[s]} n_i^p (w_i^*)^{1-p} \leq \frac{\varphi_p(\boldsymbol{n})}{\varphi_p(\boldsymbol{w}^*)}.$$

Proof. We consider the problem in which the matrix M_i is replicated n_i times:

$$\forall i \in [s],\ \forall k \in [n_i], M_{i,k} = M_i.$$

Since $\boldsymbol{w}^*$ is optimal for Problem (3), it is clear that $(w_{i,k})_{(i,k)\in\cup_{j\in[s]}\{j\}\times[n_j]}$ is optimal for the problem with replicated matrices if

$$\forall i \in [s], \sum_{k\in[n_i]} w_{i,k} = w_i^*, \tag{4}$$

i.e. $w_{i,k}$ is the part of w_i^* allocated to the k^{th} copy of the matrix M_i. For such a vector, Proposition 3.1 shows that

$$\frac{\varphi_p(\boldsymbol{n})}{\varphi_p(\boldsymbol{w}^*)} \geq \frac{1}{N}\sum_{i=1}^{s}\sum_{k=1}^{n_i}(w_{i,k}^*)^{1-p}.$$

Finally, it is easy to see (by concavity) that the latter lower bound is maximized with respect to the constraints of Equation (4) if $\forall i \in [s], \forall k \in [n_i],\ w_{i,k} = \frac{w_i^*}{n_i}$:

$$\frac{\varphi_p(\boldsymbol{n})}{\varphi_p(\boldsymbol{w^*})} \geq \frac{1}{N} \sum_{i=1}^{s} \sum_{k=1}^{n_i} \left(\frac{w_i^*}{n_i}\right)^{1-p} = \frac{1}{N} \sum_{i=1}^{s} n_i^p (w_i^*)^{1-p}.$$

□

We next give a simple rounding algorithm which finds the feasible design $\boldsymbol{n}$ which maximizes the lower bound of Proposition 3.3:

$$\max_{\substack{\boldsymbol{n}\in\mathbb{N}^s \\ \sum n_i = N}} \quad \sum_{j\in[s]} n_j^p\ w_j^{1-p}. \tag{5}$$

The latter maximization problem is in fact a *ressource allocation problem with a convex separable objective*, and the incremental algorithm which we give below is well known in the resource allocation community (see e.g. [IK88]).

Algorithm 3.1 [Incremental rounding]

```
  Input: A nonnegative vector $\boldsymbol{w} \in \mathbb{R}^s$ such that $\sum_{i=1}^s w_i = N \in \mathbb{N} \setminus \{0\}$.
  Sort the coordinates of $\boldsymbol{w}$; We assume wlog that $w_1 \geq w_2 \geq \ldots \geq w_s$;
  $\boldsymbol{n} \leftarrow [1, 0 \ldots, 0] \in \mathbb{R}^s$
  for $k = 2 \ldots N$ do
    Select an index $i_{max} \in \operatorname{argmax}_{i\in[s]} \left((n_i+1)^p - n_i^p\right)\ w_i^{1-p}$
    $n_{i_{max}} \leftarrow n_{i_{max}} + 1$
  end for
  return: a $N-$replicated design $\boldsymbol{n}$ which maximizes $\sum_{i=1}^s n_i^p w_i^{1-p}$.
```

Remark 3.4. If $\boldsymbol{w}$ is sorted ($w_1 \geq w_2 \geq \ldots \geq w_s$), then the solution of Problem (5) clearly satisfies $n_1 \geq n_2 \geq \ldots \geq n_s$. Consequently, it is not necessary to test every index $i \in [s]$ to compute the argmax in Algorithm 3.1. Instead, one only needs to compute the increments $\left((n_i+1)^p - n_i^p\right)\ w_i^{1-p}$ for the $i \in [s]$ such that $i = 1$ or $n_i + 1 \leq n_{i-1}$.

We shall now give a posterior bound for the budgeted problem (P_p^{bdg}). We only provide a sketch of the proof, since the reasoning is the same as for Propositions 3.1 and 3.3. We also point out that the approximation bound provided in the next proposition can be maximized over the set of $B-$budgeted designs, thanks to a dynamic programming algorithm which we do not detail here (see [MM76]).

Proposition 3.5. *Let $p \in [0,1]$ and $\boldsymbol{w^*}$ be optimal for the continuous relaxation*

$$\max_{\boldsymbol{w}\in\mathbb{R}^s} \left\{ \Phi_p\big(M_F(\boldsymbol{w})\big) :\ \boldsymbol{w} \geq \boldsymbol{0},\ \sum_{i\in[s]} c_i w_i \leq B \right\} \tag{6}$$

of Problem (P_p^{bdg}). Then, for any design $\boldsymbol{n} \in \mathbb{N}^s$, the following inequality holds:

$$\frac{1}{B} \sum_{i\in[s]} c_i n_i^p (w_i^*)^{1-p} \leq \frac{\varphi_p(\boldsymbol{n})}{\varphi_p(\boldsymbol{w^*})}.$$

Proof. First note that after the change of variable $z_i := NB^{-1}c_iw_i$, the continuous relaxation (6) can be rewritten under the standard form (3), where the matrix M_i is replaced by $M_i' = B(Nc_i)^{-1}M_i$. Hence, we know from Proposition B.4 that the optimality conditions of Problem (6) are:

$$\forall i \in [s], \quad Bc_i^{-1} \operatorname{trace}(M_F(\boldsymbol{w^*})^{p-1}M_i) \leq \varphi_p\big(\boldsymbol{w^*}\big),$$

with inequality if $w_i^* > 0$. Then, we can apply exactly the same reasonning as in the proof of Proposition 3.1, to show that

$$\forall S \subset [s], \quad \frac{1}{B}\sum_{i\in S} c_i(w_i^*)^{1-p} \leq \frac{\varphi_p(S)}{\varphi_p\big(\boldsymbol{w^*}\big)}.$$

The only change is that the optimality conditions must be multiplied by a factor proportional to $c_i(w_i^*)^{1-p}$ (instead of $(w_i^*)^{1-p}$ as in Equation (25)). Finally, we can apply the same arguments as in the proof of 3.3 to obtain the inequality of this proposition. □

3.3 Prior bounds

In this section, we derive *prior bounds* for the solution obtained by rounding the continuous solution of Problem (3), i.e. approximation bounds which depend only on the parameters p, N and s of Problems (P_p) and (P_p^{bin}). We first need to state one technical lemma.

Lemma 3.6. *Let $\boldsymbol{w} \in \mathbb{R}^s$ be a nonnegative vector summing to $r \leq s$, $r \in \mathbb{N}$, and p be an arbitrary real in the interval $[0,1]$. Assume without loss of generality that the coordinates of $\boldsymbol{w}$ are sorted, i.e. $w_1 \geq \ldots \geq w_s \geq 0$. If one of the following two conditions holds:*

$$\begin{aligned}(i) &\quad \forall i \in [s],\ w_i \leq 1;\\ (ii) &\quad p \leq 1 - \frac{\ln r}{\ln s},\end{aligned}$$

then, the following inequality holds:

$$\frac{1}{r}\sum_{i=1}^{r} w_i^{1-p} \geq \left(\frac{r}{s}\right)^{1-p}.$$

Proof. We start by showing the lemma under the condition (i). To this end, we consider the minimization problem

$$\min_{\boldsymbol{w}}\{\sum_{i=1}^{r} w_i^{1-p} : \sum_{i=1}^{s} w_i = r;\ 1 \geq w_1 \geq \ldots \geq w_s \geq 0\}. \tag{7}$$

Our first claim is that the optimum is necessarily attained by a vector of the form $\boldsymbol{w} = [u+\alpha_1, \ldots, u+\alpha_r, u, \ldots, u]^T$, where $\alpha_1, \ldots, \alpha_r \geq 0$, i.e. the $s-r$ coordinates of $\boldsymbol{w}$ which are not involved in the objective function are equal. To see this, assume *ad absurbium* that $\boldsymbol{w}$ is optimal for Problem (7), with $w_i > w_{i+1}$ for an index $i > r$. Define k as the smallest integer such that $w_1 = w_2 = \ldots = w_k > w_{k+1}$. Then, $\boldsymbol{e_i} - 1/k\sum_{j\in[k]}\boldsymbol{e_j}$ is a feasible direction along which the objective criterion $\sum_{i=1}^r w_i^{1-p}$ is decreasing, a contradiction. Problem (7) is hence equivalent to:

$$\min_{u,\boldsymbol{\alpha}}\{\sum_{i=1}^r (u+\alpha_i)^{1-p}:\ \sum_{i=1}^r \alpha_i = r-su;\ 0\leq u;\ 0\leq \alpha_i\leq 1-u\ (\forall i\in[r])\}. \tag{8}$$

It is known that the objective criterion of Problem (8) is Schur-concave, as a symmetric separable sum of concave functions (we refer the reader to the book of Marshall and Olkin [MO79] for details about the theory of majorization and Schur-concavity). This tells us that for all $u\in[0,\frac{r}{s}]$, the minimum with respect to $\boldsymbol{\alpha}$ is attained by

$$\boldsymbol{\alpha} = [\underbrace{1-u,\ldots,1-u}_{k\ \text{times}}, r-su-k(1-u),0,\ldots,0]^T,$$

where $k=\lfloor\frac{r-su}{1-u}\rfloor$ (for a given u, this vector majorizes all the vectors of the feasible set). Problem (8) can thus be reduced to the scalar minimization problem

$$\min_{u\in[0,\frac{r}{s}]}\ \left\lfloor\frac{r-su}{1-u}\right\rfloor + \Big(u+r-su-\left\lfloor\frac{r-su}{1-u}\right\rfloor(1-u)\Big)^{1-p} + \big(r-\left\lfloor\frac{r-su}{1-u}\right\rfloor-1\big)u^{1-p}.$$

It is not difficult to see that this function is piecewise concave, on the $r-1$ intervals of the form $u\in\left[\frac{r-(k+1)}{s-(k+1)},\frac{r-k}{s-k}\right]$, $k\in[r-1]$, corresponding to the domains where $k=\lfloor\frac{r-su}{1-u}\rfloor$ is constant. It follows that the minimum is attained for a u of the form $\frac{r-k}{s-k}$, where $k\in[r]$, and the problem reduces to

$$\min_{k\in[r]}\ k+(r-k)\left(\frac{r-k}{s-k}\right)^{1-p}.$$

Finally, one can check that the objective function of the latter problem is nondecreasing with respect to k, such that the minimum is attained for $k=0$ (which corresponds to the uniform weight vector $\boldsymbol{w}=[r/s,\ldots,r/s]^T$). This achieves the first part of this proof.

The proof of the lemma for the condition (ii) is similar. This time, we consider the minimization problem

$$\min_{\boldsymbol{w}}\{\sum_{i=1}^r w_i^{1-p}:\ \sum_{i=1}^s w_i = r;\ w_1\geq\ldots\geq w_s\geq 0\}. \tag{9}$$

Again, the optimum is attained by a vector of the form $\boldsymbol{w}=[u+\alpha_1,\ldots,u+\alpha_r,u,\ldots,u]^T$, which reduces the problem to:

$$\min_{u,\boldsymbol{\alpha}}\{\sum_{i=1}^r (u+\alpha_i)^{1-p}:\ \sum_{i=1}^r \alpha_i = r-su;\ u,\alpha_1,\ldots,\alpha_r\geq 0\}. \tag{10}$$

For a fixed u, the Schur-concavity of the objective function indicates that the minimum is attained for $\boldsymbol{\alpha}=[r-su,0,\ldots,0]^T$. Finally, Problem (10) reduces to the scalar minimization problem

$$\min_{u\in[0,\frac{r}{s}]}\ \big(u+(r-su)\big)^{1-p}+(r-1)u^{1-p},$$

where the optimum is always attained for $u = 0$ or $u = r/s$ by concavity. It now is easy to see that the inequality of the lemma is satisfied when the latter minimum is attained for $u = r/s$, i.e. if $r(\frac{r}{s})^{1-p} \leq r^{1-p}$, which is equivalent to the condition (ii) of the lemma. □

As a direct consequence of this lemma, we obtain a *prior* approximation bound for Problem (P_p^{bin}) when p is in a neighborhood of 0.

Theorem 3.7 (Approximation bound for $N-$binary designs obtained by rounding)**.** *Let $p \in [0,1]$, $N \leq s$ and $\boldsymbol{w}^*$ be a solution of the continuous optimal design problem (3). Let S be the $N-$binary design obtained by selecting the N largest coordinates of $\boldsymbol{w}^*$. If $p \leq 1 - \frac{\ln N}{\ln s}$, then we have*

$$\frac{\varphi_p(S)}{\varphi_p(S^*)} \geq \frac{\varphi_p(S)}{\varphi_p(\boldsymbol{w}^*)} \geq \left(\frac{N}{s}\right)^{1-p}.$$

Proof. This is straightforward if we combine the result of Proposition 3.1 and the one of Lemma 3.6 for $r = N$ and condition (ii). □

In the next theorem, we give an approximation factor for the design provided by Algorithm 3.1. This factor F is plotted as a function of p and the ratio $\frac{N}{s}$ on Figure 1. For every value of $\frac{N}{s}$, this theorem shows that there is a continuously increasing difficulty from the easy case ($p = 1$, where $F = 1$) to the most degenerate problem ($p = 0$, where $F = \min(\frac{N}{s}, 1 - \frac{s}{4N})$).

Theorem 3.8 (Approximation bound for $N-$replicated designs obtained by rounding)**.** *Let $p \in [0,1]$, $\boldsymbol{w}^*$ be a solution of the continuous optimal design problem (3) and $\boldsymbol{n}$ be the vector returned by Algorithm 3.1 for the input $\boldsymbol{w} = \boldsymbol{w}^*$. Then, we have*

$$\frac{\varphi_p(\boldsymbol{n})}{\varphi_p(\boldsymbol{n}^*)} \geq \frac{\varphi_p(\boldsymbol{n})}{\varphi_p(\boldsymbol{w}^*)} \geq F,$$

where F is defined by:

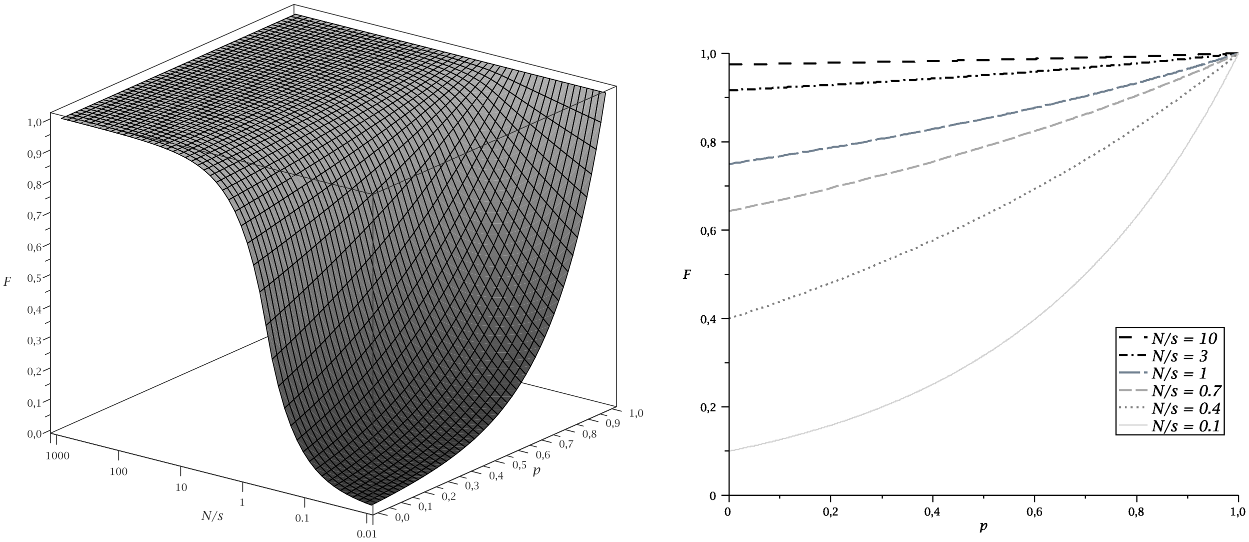
$$F = \begin{cases} \left(\frac{N}{s}\right)^{1-p} & \textit{if } \left(\frac{N}{s}\right)^{1-p} \leq \frac{1}{2-p} \quad (\textit{in particular, if } \frac{N}{s} \leq e^{-1}); \\ 1 - \frac{s}{N}(1-p)\left(\frac{1}{2-p}\right)^{\frac{2-p}{1-p}} & \textit{Otherwise} \quad (\textit{in particular, if } \frac{N}{s} \geq \frac{1}{2}); \end{cases}$$

Proof. For all $i \in [s]$ we denote by $f_i := w_i^* - \lfloor w_i^* \rfloor$ the fractional part of w_i^*, and we assume without loss of generality that these numbers are sorted, i.e. , $f_1 \geq f_2 \geq \ldots \geq f_s$. We will prove the theorem through a simple (suboptimal) rounding $\bar{\boldsymbol{n}}$, which we define as follows:

$$\bar{n}_i = \begin{cases} \lfloor w_i^* \rfloor + 1 & \text{if } i \leq N - \sum_{i \in [s]} \lfloor w_i^* \rfloor; \\ \lfloor w_i^* \rfloor & \text{Otherwise.} \end{cases}$$

We know from Proposition 3.3 and from the fact that Algorithm 3.1 solves Problem (5) the integer vector $\boldsymbol{n}$ satisfies

$$N \frac{\varphi_p(\boldsymbol{n})}{\varphi_p(\boldsymbol{w}^*)} \geq \sum_{i=1}^{s} n_i^p (w_i^*)^{1-p} \geq \sum_{i=1}^{s} \bar{n}_i^p (w_i^*)^{1-p}. \tag{11}$$



(a)

(b)

Figure 1: Approximation factor F of Theorem 3.8: (a) as a function of p and the ratio $\frac{N}{s}$ (log scale); (b) as a function of p for selected values of $\frac{N}{s}$.

We shall now bound from below the latter expression: if $\bar{n}_i = \lfloor w_i^* \rfloor$ and $\lfloor w_i^* \rfloor \neq 0$, then

$$\bar{n}_i^p (w_i^*)^{1-p} = \lfloor w_i^* \rfloor \left(\frac{w_i^*}{\lfloor w_i^* \rfloor} \right)^{1-p} \geq \lfloor w_i^* \rfloor. \tag{12}$$

Note that Inequality (12) also holds if $\bar{n}_i = \lfloor w_i^* \rfloor = 0$. If $\bar{n}_i = \lfloor w_i^* \rfloor + 1$, we write

$$\bar{n}_i^p (w_i^*)^{1-p} = \underbrace{\left(\frac{w_i^*}{\bar{n}_i} \right)^{1-p} + \ldots + \left(\frac{w_i^*}{\bar{n}_i} \right)^{1-p}}_{\bar{n}_i \text{ terms}} \geq \underbrace{1^{1-p} + \ldots + 1^{1-p}}_{\lfloor w_i^* \rfloor \text{ terms}} + f_i^{1-p} = \lfloor w_i^* \rfloor + f_i^{1-p}, \tag{13}$$

where the inequality is a consequence of the concavity of $\boldsymbol{w} \mapsto \sum_j w_j^{1-p}$. Combining Inequalities (12) and (13) yields

$$\sum_{i=1}^{s} \bar{n}_i^p (w_i^*)^{1-p} \geq \sum_{i=1}^{s} \lfloor w_i^* \rfloor + \sum_{j=1}^{N - \sum_{i=1}^{s} \lfloor w_i^* \rfloor} f_i^{1-p} = \bar{N} + \sum_{j=1}^{N-\bar{N}} f_i^{1-p},$$

where we have set $\bar{N} := \sum_{i=1}^{s} \lfloor w_i^* \rfloor \in \{\max(N-s+1, 0), \ldots, N\}$. Since the vector $\boldsymbol{f} = [f_1, \ldots, f_s]$ sums to $N - \bar{N}$, we can apply the result of Lemma 3.6 with condition (i), with $r = N - \bar{N}$, and we obtain

$$\sum_{i=1}^{s} \bar{n}_i^p w_i^{1-p} \geq \bar{N} + (N - \bar{N}) \left(\frac{N - \bar{N}}{s} \right)^{1-p} \geq \min_{u \in [0,N]} u + (N-u) \left(\frac{N-u}{s} \right)^{1-p}.$$

We will compute this lower bound in closed-form, which will provide the approximation bound of the theorem. To do this, we define the function $g : u \mapsto u + (N-u) \left(\frac{N-u}{s} \right)^{1-p}$ on

$]-\infty, N]$, and we observe (by differentiating) that g is decreasing on $]-\infty, u^*]$ and increasing on $[u^*, N[$, where

$$u^* = N - s\left(\frac{1}{2-p}\right)^{\frac{1}{1-p}}.$$

Hence, only two cases can appear: either $u^* \leq 0$, and the minimum of g over $[0, N]$ is attained for $u = 0$; or $u^* \geq 0$, and $g_{|[0,N]}$ attains its minimum at $u = u^*$. Finally, the bound given in this theorem is either $N^{-1}g(0)$ or $N^{-1}g(u^*)$, depending on the sign of u^*. In particular, since the function

$$h : p \mapsto \left(\frac{1}{2-p}\right)^{\frac{1}{1-p}}$$

is nonincreasing on the interval $[0, 1]$, with $h(0) = \frac{1}{2}$ and $h(1) = e^{-1}$, we have:

$$\forall p \in [0, 1], \quad \frac{N}{s} \leq e^{-1} \Longrightarrow u^* \leq 0 \quad \text{and} \quad \frac{N}{s} \geq \frac{1}{2} \Longrightarrow u^* \geq 0.$$

□

Remark 3.9. The alternative rounding $\tilde{\boldsymbol{n}}$ is very useful to obtain the formula of Theorem 3.8. However, since $\tilde{\boldsymbol{n}}$ differs from the design $\boldsymbol{n}$ returned by Algorithm 3.1 in general, the inequality $\frac{\varphi_p(\boldsymbol{n})}{\varphi_p(\boldsymbol{w}^*)} \geq F$ is not tight. Consider for example the situation where $p = 0$ and $N = s$, which is a trivial case for the rank optimization problem (P_0): the incremental rounding algorithm always returns a design $\boldsymbol{n}$ such that $(w_i^* > 0 \Rightarrow n_i > 0)$, and hence the problem is solve to optimality (the design is of full rank). In contrast, Theorem 3.8 only guarantees a factor $F = \frac{3}{4}$ for this class of instances.

Remark 3.10. We point out that Theorem 3.8 improves on the greedy approximation factor $1-e^{-1}$ in many situations. The gray area of Figure 2 shows the values of $(\frac{N}{s}, p) \in \mathbb{R}_+^* \times [0, 1]$ for which the approximation guarantee is better with Algorithm 3.1 than with the greedy algorithm of section 2.

Remark 3.11. Recall that the relevant criterion for the theory of optimal design is the *positively homogeneous* function $\boldsymbol{w} \mapsto \Phi_p\big(M_F(\boldsymbol{w})\big) = m^{-1/p}\varphi_p(\boldsymbol{w})^{1/p}$ (cf. Equation (20)). Hence, if a design is within a factor F of the optimum with respect to φ_p, its Φ_p-efficiency is $F^{1/p}$. In the *overinstrumented* case $N > s$, Pukelsheim gives a rounding procedure with a Φ_p-efficiency of $1 - \frac{s}{N}$ (Chapter 12 in [Puk93]). We have plotted in Figure 3 the area of the domain $(\frac{s}{N}, p) \in [0, 1]^2$ where the approximation guarantee of Theorem 3.8 is better.

4 Conclusion

This paper gives bounds on the behavior of some classical heuristics used for combinatorial problems arising in optimal experimental design. Our results can either justify or discard the use of such heuristics, depending on the settings of the instances considered. Moreover, our results confirm some facts that had been observed in the literature, namely that rounding algorithms perform better if the density of measurements is high, and that the greedy algorithm always gives a quite good solution. We illustrate these observations with two examples:

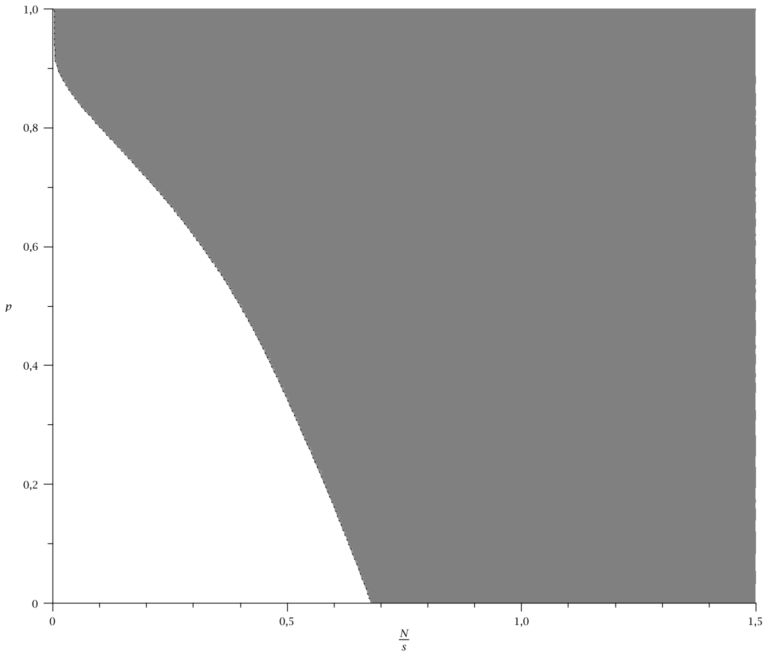


Figure 2: in gray, values of $(\frac{N}{s}, p) \in \mathbb{R}_+^* \times [0,1]$ such that the factor F of Theorem 3.8 is larger than $1 - e^{-1}$.

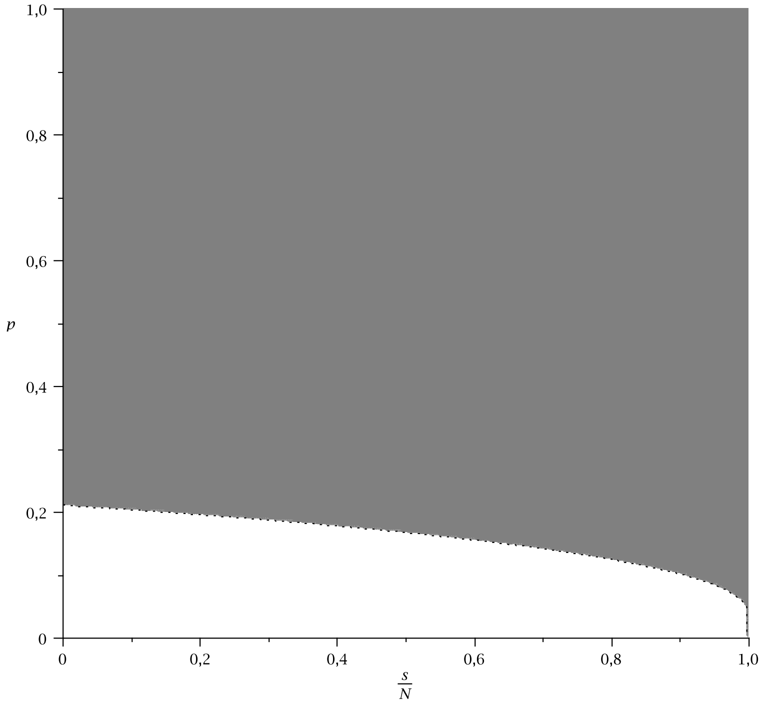


Figure 3: in gray, values of $(\frac{s}{N}, p) \in [0,1]^2$ such that the factor F of Theorem 3.8 is larger than $(1 - s/N)^p$.

In a sensor location problem, Uciński and Patan [UP07] noticed that the trimming of a Branch and Bound algorithm was better if they activated more sensors, although this led to a much larger search space. The authors claims that this surprising result can be explained by the fact that a higher density of sensors leads to a better continuous relaxation. This is confirmed by our result of approximability, which shows that the larger is the number of selected experiments, the better is the quality of the rounding.

It is also known that the greedy algorithm generally gives very good results for the optimal design of experiments (see e.g. [SQZ06], where the authors explicitly chose not to implement a local search from the design greedily chosen, since the greedy algorithm already performs very well). Our $(1-1/e)-$approximability result guarantees that this algorithm always well behaves indeed.

5 Acknowledgment

The author wishes to thank Stéphane Gaubert for his useful comments and advises, for suggesting to work within the framework of matrix inequalities and submodularity, and of course for his warm support. He also wants to thank Mustapha Bouhtou for the stimulating discussions which are at the origin of this work. The author expresses his gratitude to an anonymous referee of the ISCO conference – where an announcement of some of the present results was made [BGS10], and to three referees of DAM for precious comments and suggestions.

References

[Atw73] C.L. Atwood. Sequences converging to D-optimal designs of experiments. *Annals of statistics*, 1(2):342–352, 1973.

[AZ99] T. Ando and X. Zhan. Norm inequalities related to operator monotone functions. *Mathematische Annalen*, 315:771–780, 1999.

[BGS08] M. Bouhtou, S. Gaubert, and G. Sagnol. Optimization of network traffic measurement: a semidefinite programming approach. In *Proceedings of the International Conference on Engineering Optimization (ENGOPT)*, Rio De Janeiro, Brazil, 2008. ISBN 978-85-7650-152-7.

[BGS10] M. Bouhtou, S. Gaubert, and G. Sagnol. Submodularity and randomized rounding techniques for optimal experimental design. *Electronic Notes in Discrete Mathematics*, 36:679 – 686, March 2010. ISCO 2010 - International Symposium on Combinatorial Optimization. Hammamet, Tunisia.

[Bha97] R. Bhatia. *Matrix analysis.* Springer Verlag, 1997.

[CC84] M. Conforti and G. Cornujols. Submodular set functions, matroids and the greedy algorithm: tight worst-case bounds and some generalizations of the Rado-Edmonds theorem. *Discrete applied mathematics*, 7(3):251–274, 1984.

[CCPV07] G. Calinescu, C. Chekuri, M. Pál, and J. Vondrák. Maximizing a submodular set function subject to a matroid constraint. In *Proceedings of the 12th international conference on Integer Programming and Combinatorial Optimization, IPCO*, volume 4513, pages 182–196, 2007.

[DPZ08] H. Dette, A. Pepelyshev, and A. Zhigljavsky. Improving updating rules in multiplicative algorithms for computing D-optimal designs. *Computational Statistics & Data Analysis*, 53(2):312 – 320, 2008.

[Fed72] V.V. Fedorov. *Theory of optimal experiments*. New York : Academic Press, 1972. Translated and edited by W. J. Studden and E. M. Klimko.

[Fei98] U. Feige. A threshold of $\ln n$ for approximating set cover. *Journal of ACM*, 45(4):634–652, July 1998.

[HP95] F. Hansen and G.K. Pedersen. Perturbation formulas for traces on C*-algebras. *Publications of the research institute for mathematical sciences, Kyoto University*, 31:169–178, 1995.

[HT09] R. Harman and M. Trnovská. Approximate D-optimal designs of experiments on the convex hull of a finite set of information matrices. *Mathematica Slovaca*, 59(5):693–704, December 2009.

[IK88] T. Ibaraki and N. Katoh. *Resource allocation problems: algorithmic approaches*. MIT Press, 1988.

[JB06] E. Jorswieck and H. Boche. *Majorization and matrix-monotone functions in wireless communications*. Now Publishers Inc., 2006.

[Kie74] J. Kiefer. General equivalence theory for optimum designs (approximate theory). *The annals of Statistics*, 2(5):849–879, 1974.

[Kie75] J. Kiefer. Optimal design: Variation in structure and performance under change of criterion. *Biometrika*, 62(2):277–288, 1975.

[Kos06] T. Kosem. inequalities between $|f(a+b)|$ and $|f(a)+f(b)|$. *Linear Algebra and its Applications*, 418:153–160, 2006.

[KST09] A. Kulik, H. Shachnai, and T. Tamir. Maximizing submodular set functions subject to multiple linear constraints. In *SODA '09: Proceedings of the Nineteenth Annual ACM -SIAM Symposium on Discrete Algorithms*, pages 545–554, Philadelphia, PA, USA, 2009.

[Löw34] K. Löwner. Über monotone Matrixfunktionen. *Mathematische Zeitschrift*, 38(1):177–216, 1934.

[Min78] Michel Minoux. Accelerated greedy algorithms for maximizing submodular set functions. In J. Stoer, editor, *Optimization Techniques*, volume 7 of *Lecture Notes in Control and Information Sciences*, pages 234–243. Springer Berlin / Heidelberg, 1978. 10.1007/BFb0006528.

[MM76] T.L. Morin and R.E. Marsten. An algorithm for nonlinear knapsack problems. *Management Science*, pages 1147–1158, 1976.

[MO79] AW Marshall and I. Olkin. *Inequalities: Theory of Majorization and its Applications*. Academic Press, 1979.

[NWF78] G.L. Nemhauser, L.A. Wolsey, and M.L. Fisher. An analysis of approximations for maximizing submodular set functions. *Mathematical Programming*, 14:265–294, 1978.

[PR92] F. Pukelsheim and S. Rieder. Efficient rounding of approximate designs. *Biometrika*, pages 763–770, 1992.

[PS83] F. Pukelsheim and G.P.H. Styan. Convexity and monotonicity properties of dispersion matrices of estimators in linear models. *Scandinavian journal of statistics*, 10(2):145–149, 1983.

[Puk80] F. Pukelsheim. On linear regression designs which maximize information. *Journal of statistical planning and inferrence*, 4:339–364, 1980.

[Puk93] F. Pukelsheim. *Optimal Design of Experiments*. Wiley, 1993.

[RS89] TG Robertazzi and SC Schwartz. An Accelerated Sequential Algorithm for Producing D-Optimal Designs. *SIAM Journal on Scientific and Statistical Computing*, 10:341, 1989.

[Sag11] G. Sagnol. Computing optimal designs of multiresponse experiments reduces to second-order cone programming. *Journal of Statistical Planning and Inference*, 141(5):1684 – 1708, 2011.

[SGB10] G. Sagnol, S. Gaubert, and M. Bouhtou. Optimal monitoring on large networks by successive c-optimal designs. In *22nd international teletraffic congress (ITC22), Amsterdam, The Netherlands*, September 2010.

[SQZ06] H.H. Song, L. Qiu, and Y. Zhang. Netquest: A flexible framework for largescale network measurement. In *ACM SIGMETRICS'06*, St Malo, France, 2006.

[Svi04] M. Sviridenko. A note on maximizing a submodular set function subject to a knapsack constraint. *Operation Research Letters*, 32(1):41–43, 2004.

[UP07] D. Uciński and M. Patan. D-optimal design of a monitoring network for parameter estimation of distributed systems. *Journal of Global Optimization*, 39(2):291–322, 2007.

[Von08] J. Vondrák. Optimal approximation for the submodular welfare problem in the value oracle model. In *ACM Symposium on Theory of Computing, STOC'08*, pages 67–74, 2008.

[Wol82] L.A. Wolsey. Maximising real-valued submodular functions: Primal and dual heuristics for location problems. *Mathematics of Operations Research*, pages 410–425, 1982.

[Yu10] Y. Yu. Monotonic convergence of a general algorithm for computing optimal designs. *The Annals of Statistics*, 38(3):1593–1606, 2010.

[Zha02] X. Zhan. *Matrix Inequalities (Lecture Notes in Mathematics)*. Springer, 2002.

Appendix

A From optimal design of statistical experiments to Problem (P_p)

A.1 The classical linear model

We denote vectors by bold-face lowercase letters and we make use of the classical notation $[s] := \{1, \ldots, s\}$ (and we define $[0] := \emptyset$). The set of nonnegative (resp. positive) real numbers is denoted by $\mathbb{R}_+$ (resp. $\mathbb{R}_+^*$), the set of $m \times m$ symmetric (resp. symmetric positive semidefinite, symmetric positive definite) is denoted by $\mathbb{S}_m$ (resp. $\mathbb{S}_m^+$, $\mathbb{S}_m^{++}$). The expected value of a random variable X is denoted by $\mathbb{E}[X]$.

We denote by $\boldsymbol{\theta} \in \mathbb{R}^m$ the vector of the parameters that we want to estimate. In accordance with the classical linear model, we assume that the experimenter has a collection of s experiments at his disposal, each one providing a (multidimensional) observation which is a linear combination of the parameters, up to a noise on the measurement whose covariance matrix is known and positive definite. In other words, for each experiment $i \in [s]$, we have

$$\boldsymbol{y_i} = A_i \boldsymbol{\theta} + \boldsymbol{\epsilon_i}, \qquad \mathbb{E}[\boldsymbol{\epsilon_i}] = \mathbf{0}, \qquad \mathbb{E}[\boldsymbol{\epsilon_i}\boldsymbol{\epsilon_i}^T] = \Sigma_i, \tag{14}$$

where $\boldsymbol{y_i}$ is the vector of measurement of size l_i, A_i is a $(l_i \times m)-$matrix, and $\Sigma_i \in \mathbb{S}_{l_i}^{++}$ is a known covariance matrix. We will assume that the noises have unit variance for the sake of simplicity: $\Sigma_i = I$. We may always reduce to this case by a left multiplication of the observation equation (14) by $\Sigma_i^{-1/2}$. The errors on the measurements are assumed to be mutually independent, i.e.

$$i \neq j \Longrightarrow \mathbb{E}[\boldsymbol{\epsilon_i}\boldsymbol{\epsilon_j}^T] = 0.$$

As explained in the introduction, the aim of experimental design theory is to choose how many times each experiment will be performed so as to maximize the accuracy of the estimation of $\boldsymbol{\theta}$, with the constraint that N experiments may be conducted. We therefore define the integer-valued *design* variable $\boldsymbol{n} \in \mathbb{N}^s$, where n_k indicates how many times the experiment k is performed. We denote by $i_k \in [s]$ the index of the k^{th} conducted experiment (the order in which we consider the measurements has no importance), so that the aggregated vector of observation reads:

$$\boldsymbol{y} = \mathcal{A}\ \boldsymbol{\theta} + \boldsymbol{\epsilon}, \tag{15}$$

$$\text{where } \boldsymbol{y} = \begin{bmatrix} \boldsymbol{y_{i_1}} \\ \vdots \\ \boldsymbol{y_{i_N}} \end{bmatrix}, \qquad \mathcal{A} = \begin{bmatrix} A_{i_1} \\ \vdots \\ A_{i_N} \end{bmatrix}, \qquad \mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0}, \quad \text{and} \quad \mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T] = I.$$

Now, assume that we have enough measurements, so that $\mathcal{A}$ is of full rank. A common result in the field of statistics, known as the *Gauss-Markov* theorem, states that the best linear unbiased estimator of $\boldsymbol{\theta}$ is given by a pseudo inverse formula. Its variance is given below:

$$\hat{\boldsymbol{\theta}} = \left(\mathcal{A}^T\mathcal{A}\right)^{-1}\mathcal{A}^T\boldsymbol{y}. \tag{16}$$

$$\operatorname{Var}(\hat{\boldsymbol{\theta}}) = (\mathcal{A}^T\mathcal{A})^{-1}. \tag{17}$$

We denote the inverse of the covariance matrix (17) by $M_F(\boldsymbol{n})$, because in the Gaussian case it coincides with the Fisher information matrix of the measurements. Note that it can be decomposed as the sum of the information matrices of the selected experiments:

$$\begin{aligned} M_F(\boldsymbol{n}) &= \mathcal{A}^T\mathcal{A} \\ &= \sum_{k=1}^{N} A_{i_k}^T A_{i_k} \\ &= \sum_{i=1}^{s} n_i A_i^T A_i. \end{aligned} \tag{18}$$

The classical experimental design approach consists in choosing the design $\boldsymbol{n}$ in order to make the variance of the estimator (16) *as small as possible*. The interpretation is straightforward: with the assumption that the noise $\boldsymbol{\epsilon}$ is normally distributed, for every probability level α, the estimator $\hat{\boldsymbol{\theta}}$ lies in the confidence ellipsoid centered at $\boldsymbol{\theta}$ and defined by the following inequality:

$$(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T Q(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \leq \kappa_\alpha, \tag{19}$$

where κ_α depends on the specified probability level, and $Q = M_F(\boldsymbol{n})$ is the inverse of the covariance matrix $\text{Var}(\hat{\boldsymbol{\theta}})$. We would like to make these confidence ellipsoids *as small as possible*, in order to reduce the uncertainty on the estimation of $\boldsymbol{\theta}$. To this end, we can express the inclusion of ellipsoids in terms of matrix inequalities. The space of symmetric matrices is equipped with the *Löwner ordering*, which is defined by

$$\forall B, C \in \mathbb{S}_m, \qquad B \succeq C \Longleftrightarrow B - C \in \mathbb{S}_m^+.$$

Let $\boldsymbol{n}$ and $\boldsymbol{n'}$ denote two designs such that the matrices $M_F(\boldsymbol{n})$ and $M_F(\boldsymbol{n'})$ are invertible. One can readily check that for any value of the probability level α, the confidence ellipsoid (19) corresponding to $Q = M_F(\boldsymbol{n})$ is included in the confidence ellipsoid corresponding to $Q' = M_F(\boldsymbol{n'})$ if and only if $M_F(\boldsymbol{n}) \succeq M_F(\boldsymbol{n'})$. Hence, we will prefer the design $\boldsymbol{n}$ to the design $\boldsymbol{n'}$ if the latter inequality is satisfied.

A.2 Statement of the optimization problem

Since the Löwner ordering on symmetric matrices is only a partial ordering, the problem consisting in maximizing $M_F(\boldsymbol{n})$ is ill-posed. So we will rather maximize a scalar *information function* of the Fisher matrix, i.e. a function mapping $\mathbb{S}_m^+$ onto the real line, and which satisfies natural properties, such as positive homogeneity, monotonicity with respect to Löwner ordering, and concavity. For a more detailed description of the information functions, the reader is referred to the book of Pukelsheim [Puk93], who makes use of the class of matrix means Φ_p, as first proposed by Kiefer [Kie75]. These functions are defined like the L_p-norm of the vector of eigenvalues of the Fisher information matrix, but for $p \in [-\infty, 1]$: for a symmetric positive definite matrix $M \in \mathbb{S}_m^{++}$, Φ_p is defined by

$$\Phi_p(M) = \begin{cases} \lambda_{\min}(M) & \text{for } p = -\infty \ ; \\ (\frac{1}{m} \text{ trace } M^p)^{\frac{1}{p}} & \text{for } p \in \ ]-\infty, 1], \ p \neq 0; \\ (\det(M))^{\frac{1}{m}} & \text{for } p = 0, \end{cases} \tag{20}$$

where we have used the extended definition of powers of matrices M^p for arbitrary real parameters p: if $\lambda_1, \dots, \lambda_m$ are the eigenvalues of M counted with multiplicities, trace $M^p = \sum_{j=1}^m \lambda_j^p$. For singular positive semi-definite matrices $M \in \mathbb{S}_m^+$, Φ_p is defined by continuity:

$$\Phi_p(M) = \begin{cases} 0 & \text{for } p \in [-\infty, 0] \ ; \\ (\frac{1}{m} \text{ trace } M^p)^{\frac{1}{p}} & \text{for } p \in]0, 1]. \end{cases} \tag{21}$$

The class of functions Φ_p includes as special cases the classical optimality criteria used in the experimental design literature, namely $E-$optimality for $p = -\infty$ (smallest eigenvalue of $M_F(\boldsymbol{n})$), $D-$optimality for $p = 0$ (determinant of the information matrix), $A-$optimality for $p = -1$ (harmonic average of the eigenvalues), and $T-$optimality for $p = 1$ (trace). The case $p = 0$ (D-optimal design) admits a simple geometric interpretation: the volume of the confidence ellipsoid (19) is given by $C_m \kappa_\alpha^{m/2} \det(Q)^{-1/2}$ where $C_m > 0$ is a constant depending only on the dimension. Hence, maximizing $\Phi_0(M_F(\boldsymbol{n}))$ is the same as minimizing the volume of every confidence ellipsoid.

We can finally give a mathematical formulation to the problem of selecting N experiments to conduct among the set $[s]$:

$$\begin{aligned} \max_{n_i \in \mathbb{N} \ (i=1,\dots,s)} \quad & \Phi_p\Big(\sum_{i=1}^s n_i A_i^T A_i\Big) \\ \text{s. t.} \quad & \sum_i n_i \leq N, \end{aligned} \tag{22}$$

A.3 The underinstrumented situation

We note that the problem of maximizing the information matrix $M_F(\boldsymbol{n})$ with respect to the Löwner ordering remains meaningful even when $M_F(\boldsymbol{n})$ is not of full rank (the interpretation of $M_F(\boldsymbol{n})$ as *the inverse of the covariance matrix of the best linear unbiased estimator* vanishes, but $M_F(\boldsymbol{n})$ is still the Fisher information matrix of the experiments if the measurement errors are Gaussian). This case does arise in *underinstrumented situations*, in which some constraints may not allow one to conduct a number of experiments which is sufficient to infer all the parameters.

An interesting and natural idea to find an optimal under-instrumented design is to choose the design which maximizes the rank of the observation matrix $\mathcal{A}$, or equivalently of $M_F(\boldsymbol{n}) = \mathcal{A}^T \mathcal{A}$. The *rank maximization* is a nice combinatorial problem, where we are looking for a subset of matrices whose sum is of maximal rank:

$$\begin{aligned} \max_{\boldsymbol{n} \in \mathbb{N}^s} \quad & \text{rank} \Big(\sum_i n_i A_i^T A_i\Big) \\ \text{s. t.} \quad & \sum_i n_i \leq N. \end{aligned}$$

When every feasible information matrix is singular, Equation (21) indicates that the maximization of $\Phi_p(M_F(\boldsymbol{n}))$ can be considered only for nonnegative values of p. Then, the next proposition shows that Φ_p can be seen as a deformation of the rank criterion for $p \in]0,1]$. First notice that when $p > 0$, the maximization of $\Phi_p(M_F(\boldsymbol{n}))$ is equivalent to:

$$\begin{aligned} \max_{\boldsymbol{n}\in\mathbb{N}^s} \quad & \varphi_p(\boldsymbol{n}) := \operatorname{trace}\Big(\sum_i n_i A_i^T A_i\Big)^p \\ \text{s. t.} \quad & \sum_i n_i \leq N. \end{aligned}$$

If we set $M_i = A_i^T A_i$, we obtain the problems (P_0) and (P_p) which were presented in the first lines of this article.

Proposition A.1. *For all positive semidefinite matrix $M \in \mathbb{S}_m^+$,*

$$\lim_{p\to 0^+} \operatorname{trace} M^p = \operatorname{rank} M. \tag{23}$$

Proof. Let $\lambda_1, \ldots, \lambda_r$ denote the positive eigenvalues of M, counted with multiplicities, so that r is the rank of M. We have the first order expansion as $p \to 0^+$:

$$\operatorname{trace} M^p = \sum_{k=1}^r \lambda_k^p = r + p \;\log(\prod_{k=1}^r \lambda_k) + \mathcal{O}(p^2) \tag{24}$$

□

Consequently, trace M^0 will stand for rank(M) in the sequel and the rank maximization problem (P_0) is the limit of problem (P_p) as $p \to 0^+$.

Corollary A.2. *If $p > 0$ is small enough, then every design $\boldsymbol{n}^*$ which is a solution of Problem (P_p) maximizes the rank of $M_F(\boldsymbol{n})$. Moreover, among the designs which maximize this rank, $\boldsymbol{n}^*$ maximizes the product of nonzero eigenvalues of $M_F(\boldsymbol{n})$.*

Proof. Since there is only a finite number of designs, it follows from (24) that for $p > 0$ small enough, every design which maximizes φ_p must maximize in the lexicographical order first the rank of $M_F(\boldsymbol{n})$, and then the pseudo-determinant $\prod_{\{k:\lambda_k>0\}} \lambda_k$. □

B Proof of Proposition 3.1

The proof of Proposition 3.1 relies on several lemmas on the directional derivative of a scalar function applied to a symmetric matrix, which we state next. First recall that if f is differentiable on $\mathbb{R}_+^*$, then f is Fréchet differentiable over $\mathbb{S}_m^{++}$, and for $M \in \mathbb{S}_m^{++}$, $H \in \mathbb{S}_m$, we denote by $Df(M)(H)$ its directional derivative at M in the direction of H (see Equation (2)).

Lemma B.1. *If f is continuously differentiable on $\mathbb{R}_+^*$, i.e. $f \in \mathcal{C}^1(\mathbb{R}_+^*)$, $M \in \mathbb{S}_m^{++}$, $A, B \in \mathbb{S}_m$, then*

$$\operatorname{trace}(A \; Df(M)(B)) = \operatorname{trace}(B \; Df(M)(A)).$$

Proof. Let $M = QDQ^T$ be an eigenvalue decomposition of M. It is known (see e.g. [Bha97]) that $Df(M)(H)$ can be expressed as $Q(f^{[1]}(D) \odot Q^T HQ)Q^T$, where $f^{[1]}(D)$ is a symmetric matrix called the *first divided difference* of f at D and $\odot$ denotes the Hadamard (elementwise) product of matrices. With little work, the latter derivative may be rewritten as:

$$Df(M)(H) = \sum_{i,j} f_{ij}^{[1]} \boldsymbol{q_i}\boldsymbol{q_i}^T H \boldsymbol{q_j}\boldsymbol{q_j}^T,$$

where $\boldsymbol{q_k}$ is the k^{th} eigenvector of M (i.e., the k^{th} column of Q) and $f_{ij}^{[1]}$ denotes the $(i,j)-$element of $f^{[1]}(D)$. We can now conclude:

$$\begin{aligned} \operatorname{trace}(A\ Df(M)(B)) &= \sum_{i,j} f_{ij}^{[1]} \operatorname{trace}(A\boldsymbol{q_i}\boldsymbol{q_i}^T B\boldsymbol{q_j}\boldsymbol{q_j}^T) \\ &= \sum_{i,j} f_{ji}^{[1]} \operatorname{trace}(B\boldsymbol{q_j}\boldsymbol{q_j}^T H\boldsymbol{q_i}\boldsymbol{q_i}^T) \\ &= \operatorname{trace}(B\ Df(M)(A)) \end{aligned}$$

□

We next show that when f is antitone, the mapping $X \mapsto Df(M)(X)$ is nonincreasing with respect to the Löwner ordering.

Lemma B.2. *If f is differentiable and antitone on $\mathbb{R}_+^*$, then for all A, B in $\mathbb{S}_m$,*

$$A \preceq B \Longrightarrow Df(M)(A) \succeq Df(M)(B).$$

Proof. The lemma trivially follows from the definition of the directional derivative:

$$Df(M)(A) = \lim_{\epsilon \to 0^+} \frac{1}{\epsilon}\big(f(M + \epsilon A) - f(M)\big)$$

and the fact that $A \preceq B$ implies $M + \epsilon A \preceq M + \epsilon B$ for all $\epsilon > 0$. □

Lemma B.3. *Let f be differentiable on $\mathbb{R}_+^*$, $M \in \mathbb{S}_m^{++}$, $A \in \mathbb{S}_m$. If A and M commute, then*

$$Df(M)(A) = f'(M)A \in \mathbb{S}_m,$$

where f' denotes the (scalar) derivative of f.

Proof. Since A and M commute, we can diagonalize them simultaneously:

$$M = Q\operatorname{Diag}(\boldsymbol{\lambda})Q^T, \quad A = Q\operatorname{Diag}(\boldsymbol{\mu})Q^T.$$

Thus, it is clear from the definition of the directional derivative that

$$Df(M)(A) = Q\ Df\big(\operatorname{Diag}(\boldsymbol{\lambda})\big)\big(\operatorname{Diag}(\boldsymbol{\mu})\big)\ Q^T.$$

By reasoning entry-wise on the diagonal matrices, we find:

$$Df\big(\operatorname{Diag}(\boldsymbol{\lambda})\big)\big(\operatorname{Diag}(\boldsymbol{\mu})\big) = \operatorname{Diag}\big(f'(\lambda_1)\mu_1, \ldots, f'(\lambda_m)\mu_m\big) = \operatorname{Diag}\big(f'(\boldsymbol{\lambda})\big)\operatorname{Diag}(\boldsymbol{\mu})$$

The equality of the lemma is finally obtained by writing:

$$Df(M)(A) = Q\operatorname{Diag}\big(f'(\boldsymbol{\lambda})\big)\operatorname{Diag}(\boldsymbol{\mu})Q^T = Q\operatorname{Diag}\big(f'(\boldsymbol{\lambda})\big)Q^T Q\operatorname{Diag}(\boldsymbol{\mu})Q^T = f'(M)A.$$

Note that the matrix $f'(M)A$ is indeed symmetric, because $f'(M)$ and A commute. □

Before we give the proof of the main result, we recall an important result from the theory of optimal experimental designs, which characterizes the optimum of Problem (3).

Proposition B.4 (General equivalence theorem [Kie74]). *Let $p \in [0,1]$. A design $\boldsymbol{w}^*$ is optimal for Problem (3) if and only if:*

$$\forall i \in [s], \quad N \operatorname{trace}(M_F(\boldsymbol{w}^*)^{p-1} M_i) \leq \varphi_p(\boldsymbol{w}^*).$$

Moreover, the latter inequalities become equalities for all i such that $w_i^* > 0$.

For a proof of this result, see [Kie74] or Paragraph 7.19 in [Puk93], where the problem is studied with the normalized constraint $\sum_i w_i \leq 1$. In fact, the *general equivalence theorem* details the Karush-Kuhn-Tucker conditions of optimality of Problem (3). To derive them, one can use the fact that when $M_F(\boldsymbol{w})$ is invertible,

$$\frac{\partial \varphi_p(\boldsymbol{w})}{\partial w_i} = \operatorname{trace}(M_F(\boldsymbol{w})^{p-1} M_i) \quad \text{for all} \quad p \in]0,1],$$

and

$$\frac{\partial \log \det(M_F(\boldsymbol{w}))}{\partial w_i} = \operatorname{trace}(M_F(\boldsymbol{w})^{-1} M_i).$$

Note that for $p \neq 1$, the proposition implicitly implies that $M_F(\boldsymbol{w}^*)$ is invertible. A proof of this fact can be found in Paragraph 7.13 of [Puk93].

We can finally prove the main result:

Proof of Proposition 3.1. Let $\boldsymbol{w}^*$ be an optimal solution to Problem (3) and S be a subset of $[s]$ such that $w_i^* > 0$ for all $i \in S$ (the case in which $w_i^* = 0$ for some index $i \in S$ will trivially follow if we adopt the convention $0^0 = 0$). We know from Proposition B.4 that $N^{-1}\varphi_p(\boldsymbol{w}^*) = \operatorname{trace}(M_F(\boldsymbol{w}^*)^{p-1} M_i)$ for all i in S. If we combine these equalities by multiplying each expression by a factor proportional to $(w_i^*)^{1-p}$, we obtain:

$$\begin{aligned} \frac{1}{N}\varphi_p(\boldsymbol{w}^*) &= \sum_{i \in S} \frac{(w_i^*)^{1-p}}{\sum_{k \in S}(w_k^*)^{1-p}} \operatorname{trace}(M_F(\boldsymbol{w}^*)^{p-1} M_i) \qquad (25) \\ \iff \frac{1}{N} \sum_{k \in S} (w_k^*)^{1-p} &= \frac{\sum_{i \in S}(w_i^*)^{1-p} \operatorname{trace}(M_F(\boldsymbol{w}^*)^{p-1} M_i)}{\varphi_p(\boldsymbol{w}^*)}. \end{aligned}$$

We are going to show that for all $\boldsymbol{w} \geq \boldsymbol{0}$ such that $M_F(\boldsymbol{w})$ is invertible, $\sum_{i \in S} w_i^{1-p} \operatorname{trace}(M_F(\boldsymbol{w})^{p-1} M_i) \leq \operatorname{trace}(M_S)^p$, where $M_S := \sum_{i \in S} M_i$, which will complete the proof. To do this, we introduce the function f defined on the open subset of $(\mathbb{R}_+)^s$ such that $M_F(\boldsymbol{w})$ is invertible by:

$$f(\boldsymbol{w}) = \sum_{i \in S} w_i^{1-p} \operatorname{trace}(M_F(\boldsymbol{w})^{p-1} M_i) = \operatorname{trace}\left(\Big(\sum_{i \in S} w_i^{1-p} M_i\Big) M_F(\boldsymbol{w})^{p-1}\right).$$

Note that f satisfies the property $f(t\boldsymbol{w}) = f(\boldsymbol{w})$ for all positive scalar t; this explains why we do not have to work with normalized designs such that $\sum_i w_i = N$. Now, let $\boldsymbol{w} \geq \boldsymbol{0}$ be such that $M_F(\boldsymbol{w}) \succ 0$ and let k be an index of S such that $w_k = \min_{i \in S} w_i$. We are first going to show that $\frac{\partial f(\boldsymbol{w})}{\partial w_k} \geq 0$. By the rule of differentiation of a product,

$$\begin{aligned}
\frac{\partial f(\boldsymbol{w})}{\partial w_k} &= \operatorname{trace}\left((1-p)w_k^{-p} M_k M_F(\boldsymbol{w})^{p-1} + \Big(\sum_{i \in S} w_i^{1-p} M_i\Big) \frac{\partial (M_F(\boldsymbol{w})^{p-1})}{\partial w_k}\right) \\
&= \operatorname{trace}\left((1-p)w_k^{-p} M_k M_F(\boldsymbol{w})^{p-1} + \Big(\sum_{i \in S} w_i^{1-p} M_i\Big) D[x \mapsto x^{p-1}](M_F(\boldsymbol{w}))(M_k)\right) \\
&= \operatorname{trace} M_k \left((1-p)w_k^{-p} M_F(\boldsymbol{w})^{p-1} + D[x \mapsto x^{p-1}]\big(M_F(\boldsymbol{w})\big)\big(\sum_{i \in S} w_i^{1-p} M_i\big)\right), \qquad (26)
\end{aligned}$$

where the first equality is simply a rewriting of $\frac{\partial (M_F(\boldsymbol{w})^{p-1})}{\partial w_k}$ by using a directional derivative, and the second equality follows from Lemma B.1 applied to the function $x \mapsto x^{p-1}$. By linearity of the Fréchet derivative, we have:

$$w_k^p\, D[x \mapsto x^{p-1}]\big(M_F(\boldsymbol{w})\big)\big(\sum_{i \in S} w_i^{1-p} M_i\big) = D[x \mapsto x^{p-1}]\big(M_F(\boldsymbol{w})\big)\big(\sum_{i \in S} w_i \left(\frac{w_k}{w_i}\right)^p M_i\big).$$

Since $w_k \leq w_i$ for all $i \in S$, the following matrix inequality holds:

$$\sum_{i \in S} w_i \left(\frac{w_k}{w_i}\right)^p M_i \preceq \sum_{i \in S} w_i M_i \preceq M_F(\boldsymbol{w}).$$

By applying successively Lemma B.2 ($x \mapsto x^{p-1}$ is antitone on $\mathbb{R}_+^*$) and Lemma B.3 (the matrix $M_F(\boldsymbol{w})$ commutes with itself), we obtain:

$$\begin{aligned}
w_k^p\, D[x \mapsto x^{p-1}]\big(M_F(\boldsymbol{w})\big)\big(\sum_{i \in S} w_i^{1-p} M_i\big) &\succeq D[x \mapsto x^{p-1}]\big(M_F(\boldsymbol{w})\big)\big(M_F(\boldsymbol{w})\big) \\
&= (p-1) M_F(\boldsymbol{w})^{p-2} M_F(\boldsymbol{w}) \\
&= (p-1) M_F(\boldsymbol{w})^{p-1}.
\end{aligned}$$

Dividing the previous matrix inequality by w_k^p, we find that the matrix that is inside the largest parenthesis of Equation (26) is positive semidefinite, from which we can conclude: $\frac{\partial f(\boldsymbol{w})}{\partial w_k} \geq 0$.

Thanks to this property, we next show that $f(\boldsymbol{w}) \leq f(\boldsymbol{v})$, where $\boldsymbol{v} \in \mathbb{R}^s$ is defined by $v_i = \max_{k \in S}(w_k)$ if $i \in S$ and $v_i = w_i$ otherwise. Assume without loss of generality (after a reordering of the coordinates) that $S = [s_0]$, $w_1 \leq w_2 \leq \ldots \leq w_{s_0}$, and denote the vector of the remaining components of $\boldsymbol{w}$ by $\bar{\boldsymbol{w}}$ (i.e., we have $\boldsymbol{w}^T = [w_1, \ldots, w_{s_0}, \bar{\boldsymbol{w}}]$ and $\boldsymbol{v}^T = [w_{s_0}, \ldots, w_{s_0}, \bar{\boldsymbol{w}}]$). The following inequalities hold:

$$f(\boldsymbol{w}) = f\left(\begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ \vdots \\ w_{s_0} \\ \bar{\boldsymbol{w}} \end{bmatrix}\right) \leq f\left(\begin{bmatrix} w_2 \\ w_2 \\ w_3 \\ \vdots \\ w_{s_0} \\ \bar{\boldsymbol{w}} \end{bmatrix}\right) \leq f\left(\begin{bmatrix} w_3 \\ w_3 \\ w_3 \\ \vdots \\ w_{s_0} \\ \bar{\boldsymbol{w}} \end{bmatrix}\right) \leq \ldots \leq f\left(\begin{bmatrix} w_{s_0} \\ w_{s_0} \\ w_{s_0} \\ \vdots \\ w_{s_0} \\ \bar{\boldsymbol{w}} \end{bmatrix}\right) = f(\boldsymbol{v}).$$

The first inequality holds because $\frac{\partial f(\boldsymbol{w})}{\partial w_1} \geq 0$ as long as $w_1 \leq w_2$. To see that the second inequality holds, we apply the same reasoning on the function $\tilde{f} : [w_2, w_3, \ldots] \mapsto f([w_2, w_2, w_3, \ldots])$, i.e., we consider a variant of the problem where the matrices M_1 and M_2 have been replaced by a single matrix $M_1 + M_2$. The following inequalities are obtained in a similar manner.

Recall that we have set $M_S = \sum_{i \in S} M_i$. We have:

$$M_F(\boldsymbol{v}) = w_{s_0} M_S + \sum_{i \notin S} w_i M_i \succeq w_{s_0} M_S$$

and by isotonicity of the mapping $x \mapsto x^{1-p}$, $M_F(\boldsymbol{v})^{1-p} \succeq (w_{s_0}\ M_S)^{1-p}$.

We denote by $X^\dagger$ the Moore-Penrose inverse of X. It is known [PS83] that if $M_i \in \mathbb{S}_m^+$, the function $X \mapsto \operatorname{trace}(X^\dagger M_i)$ is nondecreasing with respect to the Löwner ordering over the set of matrices X whose range contains M_i. Hence, since $M_F(\boldsymbol{v}) \succeq M_F(\boldsymbol{w})$ is invertible,

$$\forall i \in S, \quad \operatorname{trace}(M_F(\boldsymbol{v})^{p-1} M_i) = \operatorname{trace}\left(\left(M_F(\boldsymbol{v})^{1-p}\right)^\dagger M_i\right) \leq \operatorname{trace}\left(\left((w_{s_0}\ M_S)^{1-p}\right)^\dagger M_i\right)$$

and

$$\begin{aligned} f(\boldsymbol{v}) &= w_{s_0}^{1-p} \sum_{i \in S} \operatorname{trace}(M_F(\boldsymbol{v})^{p-1} M_i) \\ &\leq w_{s_0}^{1-p} \sum_{i \in S} \operatorname{trace}\left(\left((w_{s_0}\ M_S)^{1-p}\right)^\dagger M_i\right) \\ &= \operatorname{trace}\left(M_S^{1-p}\right)^\dagger M_S \\ &= \operatorname{trace} M_S^p \end{aligned}$$

Finally, we have $f(\boldsymbol{w}) \leq f(\boldsymbol{v}) \leq \operatorname{trace} M_S^p = \varphi_p(S)$, and the proof is complete. □